Zinc and the common cold: problems in the review by Caruso et al. (2007)

Harri Hemilä
Department of Public Health
University of Helsinki,
Helsinki, Finland
harri.hemila@helsinki.fi
http://www.mv.helsinki.fi/home/hemila
version 10 Sept, 2013

This is an unpublished commentary on the following paper:

Caruso TJ, Prober CG, Gwaltney JM
Treatment of naturally acquired common colds with zinc: a structured review
Clin Infect Dis **2007**;45:569-74
http://dx.doi.org/10.1086/520031
http://www.ncbi.nlm.nih.gov/pubmed/17682990

These comments were motivated by differences between the above analysis and the following meta-analysis published in 2011:

Hemilä H
Zinc lozenges may shorten the duration of colds: a systematic review.
Open Respir Med J **2011**;5:51-8.
http://www.ncbi.nlm.nih.gov/pubmed/21769305
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3136969
See references available for this paper as links at:
http://www.mv.helsinki.fi/home/hemila/Zn/TORMJ.htm

In their systematic review, Caruso et al. identified 14 zinc trials (the above paper). They used the quality scoring approach so that for the identified trials they gave one point for each of 11 quality items if it was satisfied. In two tables and one figure, Caruso et al. described the distribution of quality scores and the individual quality features of the trials. They considered that only studies with the full 11 points were valid: "*Four studies met all 11 criteria. Three of these studies reported no therapeutic effect from zinc lozenge or nasal spray. One study reported positive results from zinc nasal spray.*" On the basis of this 3 vs. 1 comparison (so called "vote counting" [1]) Caruso et al. concluded that "*the therapeutic effectiveness of zinc lozenges has yet to be established.*" They proposed that the positive findings with zinc could be explained by methodological faults in the trials.

The approach to evaluate trial quality by a set of explicit criteria was initiated by Thomas Chalmers in the early 1980s [2]. Thereafter dozens of quality scales have been developed. However, the approach was not successful and it is discouraged for example in the Cochrane Handbook (2011), which states that "*The use of scales for assessing quality or risk of bias is explicitly discouraged in Cochrane reviews.*" [3].

One major problem of quality scoring is the focus on reporting in contrast to the scientific quality of the trial. For example, Caruso et al. give one point if there was "measurement of dropout rate" in the trial. This means that a trial can report high dropout rate, which means low scientific quality, yet the trial gets one point from Caruso et al., because the high dropout rate was reported explicitly. Caruso et al. give one point for "sample size calculation" which is important when a trial is planned, because it can show that the planned trial is too small, whereas it is irrelevant after the trial is published, because then the confidence interval reveals the accuracy of the result. Most of Caruso et al.'s remaining nine quality items have similar problems, see below for detailed comments.

Although it is important to consider the methods of a trial, there are no simple criteria which describe whether a trial is reliable or not. For example, in a meta-analysis of 276 randomized controlled trials, Balk et al. concluded that "*double blinding and allocation concealment, two quality measures that are frequently used in meta-analyses, were not associated with treatment effect*" [4] meaning that valid estimates of treatment effect can be reached without them. Furthermore, Glasziou et al. pointed out that in some cases firm conclusions of treatment benefit can be drawn even without any control groups [5].

Furthermore, Caruso et al. did not present the numerical results of the trials, simply classifying them as positive (statistically significant effect) or negative (no statistically significant effect), even though such a "vote counting" approach has been strongly discouraged [1]. For example, a large number of placebo-controlled trials on vitamin C and common cold found non-significant effects on common cold duration, but the results consistently favoured vitamin C, and quantitative pooling of the results showed a statistically highly significant benefit from the vitamin [6,7].

Caruso et al. did not discuss the possibility that the dose of zinc or the lozenge composition might have an effect on trial results, nor did they refer to any of the numerous papers that discussed the possibility that the level of free zinc ion might be an important variable in zinc lozenge trials [8-13].

While Caruso et al. focused on the methodological features, mostly irrelevant to trial validity, they stated that a "*common deficiency [in the zinc trials] was proof of blinding which was lacking in 7 studies. The placebo effect in the treatment of colds was first shown >70 years ago and has since been demonstrated in subsequent studies*".
As a justification for this statement, Caruso et al. referred to the Chalmers review [14] and the Karlowski trial [15].

However, Caruse et al. knew that those two papers [14,15] were erroneous, because that was pointed

out to them in a criticism of their earlier biased review on echinacea and the common cold [16,17]: "*The Chalmers review* [14] *was shown to be erroneous a decade ago; it has data inconsistent with the original study publications, errors in calculations, and other problems*".

The Karlowski trial found statistically significant benefit of vitamin C against the common cold, yet Karlowski et al. concluded paradoxically that "*the effects [of vitamin C] demonstrated might be explained equally well by a break in the double blind*" [15, p. 1038].

The comments on the Caruso et al. review on echinacea and the common cold described the problems of the Karlowski trial as follows [17]:
"*the [Karlowski] subgroup analysis excluded 105 episodes of common cold (42% of all episodes of cold), even though the 2 subgroups were presented as if they were complementary. There are numerous additional problems with Karlowski's placebo effect explanation, and, consequently, it is not a valid interpretation to the study results.*"

The previous criticism [17] referred to papers which explicitly documented in detail the problems of the Chalmers review [18] and the Karlowski trial [19].
Thus, in their zinc review in 2007, Caruso et al. kept referring to those old erroneous papers [14,15] although they knew them to be erroneous.

See specific comments on the Caruso et al. 11 point scoring system after the references.

References:

1 Rothman KJ, Greenland S (1998) Meta-analysis. Some methods to avoid: Qualitative tally (vote counting) and Quality scoring. In: Modern Epidemiology 2nd ed. New York: Lippincott Williams Wilkins. p. 671-2.

2 Chalmers TC, et al. (1981) A method for assessing the quality of a randomized control trial. Control Clin Trials 2: 31-49
http://dx.doi.org/10.1016/0197-2456(81)90056-8

3 Higgins JPT, Green S (eds.) (2011) Cochrane Handbook for Systematic Reviews of Interventions. Ver 5.1. [March 2011]. Section 8.3.3. The Cochrane Collaboration.
http://www.cochrane-handbook.org

4 Balk EM, et al. (2002) Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 287: 2973-82
http://dx.doi.org/10.1001/jama.287.22.2973

5 Glasziou P, Chalmers I, Rawlins M, McCulloch P (2007)
When are randomised trials unnecessary? Picking signal from noise. BMJ 334: 349-51
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1800999

6 Hemilä H, Chalker EB (2013) Vitamin C for preventing and treating the common cold. Cochrane Database Syst Rev CD000980.
http://dx.doi.org/10.1002/14651858.CD000980.pub4
http://www.mv.helsinki.fi/home/hemila/CC/CC.htm

7 Douglas RM, Hemilä H (2005) Vitamin C for preventing and treating the common cold. PLoS Medicine 2: e168
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160577

8 Godfrey JC (1988) Zinc for the common cold. Antimicrob Agents Chemother 32: 605-6
http://aac.asm.org/cgi/reprint/32/4/605

9 Eby GA (1988) Stability constants of zinc complexes affect common cold treatment results. Antimicrob Agents Chemother 32: 606-7
http://aac.asm.org/cgi/reprint/32/4/606

10 Martin RB (1988) pH as a variable in free zinc ion concentration from zinc-containing lozenges. Antimicrob Agents Chemother 32: 608-9
http://aac.asm.org/cgi/reprint/32/4/608

11 Eby GA (1997) Zinc ion availability - the determinant of efficacy in zinc lozenge treatment of common colds. J Antimicrob Chemother 40: 483-93
http://jac.oxfordjournals.org/cgi/content/abstract/40/4/483
http://coldcure.com/html/jac-common-cold.html

12 Bakar NKA, Taylor DM, Williams DR (1999) The chemical speciation of zinc in human saliva: possible correlation with reduction of the symptoms of the common cold produced by zinc gluconate-containing lozenges. Chemical Speciation and Bioavailability 11: 95-101
http://dx.doi.org/10.3184/095422999782775672

13 Eby GA (2004) Zinc lozenges: cold cure or candy? Solution chemistry determinations. Biosci Rep 24: 23-39
http://dx.doi.org/10.1023/B:BIRE.0000037754.71063.41
http://george-eby-research.com/html/common-cold.pdf

14 Chalmers TC (1975) Effects of ascorbic acid on the common cold: an evaluation of the evidence. Am J Med 58: 532-6
http://dx.doi.org/10.1016/0002-9343(75)90127-8

15 Karlowski TR, Chalmers TC, Frenkel LD, Kapikian AZ, Lewis TL, Lynch JM (1975) Ascorbic acid for the common cold. JAMA 231: 1038-42
http://jama.ama-assn.org/content/231/10/1038

16 Caruso TJ, Gwaltney JM (2005) Treatment of the common cold with echinacea: a structured review. Clin Infect Dis 40: 807-10
http://dx.doi.org/10.1086/428061

17 Hemilä H (2005) Echinacea, vitamin C, the common cold, and blinding. Clin Infect Dis 41: 762-3
http:/dx.doi.org/10.1086/432629

18 Hemilä H, Herman ZS (1995) Vitamin C and the common cold: a retrospective analysis of Chalmers' review. J Am Coll Nutr 14: 116-23
http://www.ncbi.nlm.nih.gov/pubmed/7790685
http://www.mv.helsinki.fi/home/hemila/H/HH_1995.pdf

18b Hemilä. Chalmers' meta-analysis (1975)
http://www.mv.helsinki.fi/home/hemila/reviews/chalmers

19 Hemilä H (1996) Vitamin C, the placebo effect, and the common cold: a case study of how preconceptions influence the analysis of results [Discussion in 49: 1085-7]. J Clin Epidemiol 49: 1079-84
http://www.ncbi.nlm.nih.gov/pubmed/8826986
http://dx.doi.org/10.1016/0895-4356(96)00189-8
http://hdl.handle.net/10250/8082
http://www.mv.helsinki.fi/home/hemila/H/HH_1996_JCE.pdf
Replies to Chalmers comments:
http://dx.doi.org/10.1016/0895-4356(96)00191-6
http://hdl.handle.net/10250/8079
http://www.mv.helsinki.fi/home/hemila/H/HH_1996_JCE2.pdf

19b Hemilä H. The most influential trial on vitamin C and the common cold: Karlowski et al. (1975)
http://www.mv.helsinki.fi/home/hemila/karlowski

Problems with the quality scoring approach by Caruso et al. (2007)
http://dx.doi.org/10.1086/520031

Caruso et al. assessed 11 quality items, and gave one point for each of the 11 items if it was satisfied. Thus, the maximum score was 11 points. Caruso et al. considered that only those trials that received the total of 11 points were valid. Four studies got 11 points, and three of them were "negative" and one was "positive". Based on this vote counting [1], Caruso et al. concluded that "*the therapeutic effectiveness of zinc lozenges has yet to be established*."

Here below are a comment about "quality scales" by a few authorities.
The following pages briefly show the specific problems of the quality items by Caruso et al. (2007).

Cochrane Handbook version 5.1 (2011 [ref. 3]) section 8.3.3
http://handbook.cochrane.org
commented:
"The use of scales for assessing quality or risk of bias is explicitly discouraged in Cochrane reviews. While the approach offers appealing simplicity, it is not supported by empirical evidence (Emerson 1990, Schulz 1995b). Calculating a summary score inevitably involves assigning 'weights' to different items in the scale, and it is difficult to justify the weights assigned. Furthermore, scales have been shown to be unreliable assessments of validity (Jüni 1999) and they are less likely to be transparent to users of the review. It is preferable to use simple approaches for assessing validity that can be fully reported (i.e. how each trial was rated on each criterion)."

Feinstein (J Clin Epidemiol 1995; 48:71-9, p. 72)
Meta-analysis: Statistical alchemy for the 21st century
http://dx.doi.org/10.1016/0895-4356(94)00110-C
commented:
"[quality scoring] gives credit for the availability, but not the scientific quality, of the basic evidence. The original investigators may get good scores for telling us what they did, but no appraisal is given to how well they did it.
For example, if the original investigators stated criteria for the diagnosis of congestive heart failure the trial is given a positive score for the available evidence. On the other hand, if congestive heart failure is defined as 'use of digitalis', the evidence has poor scientific quality and is clinically silly, but still gets credited for being available...
The criteria of "quality" are usually aimed at the availability, not the actual quality of the evidence."

Greenland (Modern Epidemiology 2nd edn [ref. 1], 1998, p. 672) comments under title:
"Some methods to avoid: Quality scoring"
"Quality scoring submerges important information by combining disparate study features into as single score. It also introduces an unnecessary and somewhat arbitrary subjective element into the analysis via the scoring scheme. Quality scoring can and should be replaced by direct categorical and regression analyses of the impact of each quality item. Such item-specific analyses let the data, rather than the investigator, indicate the importance of each item in determining the estimated effect."

Specific problems with the 11 quality items used by Caruso et al. (2007):
http://dx.doi.org/10.1086/520031

"Validated case definition"
This is related to applicability and not validity.
If a study is methodologically valid (e.g. double-blind), the difference between the study groups is reliable but the extrapolation of findings depends on the case definition.

"Quantifiable hypothesis" and
"Sample-size calculation"
These are related to precision and not validity.
The purpose of meta-analysis is to pool all methodologically valid studies on the given topic.
Small trials should not be excluded, they simply get lower weight in pooling.
Small size *per se* does not mean that the study results are not valid.

"Randomized assignment" and
"Double blinding" and
"Proof of blinding"
As mentioned above, a meta-analyses of 276 RCTs found out that "*double blinding and allocation concealment, two quality measures that are frequently used in meta-analyses, were not associated with treatment effect*" [4].
Therefore, a study should not be considered lacking any validity if participants are not allocated randomly, but by alternative allocation, or if there is not double blinding but only single blinding, etc.

"Measurement of compliance" and
"Measurement of drop-out rate"
The validity of the trial depends on the *degree* of compliance and drop-out rate, and their possible difference between the study groups, and not on whether they are merely reported or not.
Thus, according to Caruso et al. reasoning, if the drop-out rate is reported as 50%, the trial is given a point for the explicit reporting of the drop-out rate. Another trial would not be given a point, even though it might be implicitly obvious that no one dropped out.
In small and short studies it is often implicitly evident that compliance is high and drop-out rate is low, but these issues may remain unreported, for example, because journals have strict space limitations. Thus, the relevant question is not, whether drop-out rate or compliance level are reported, but what are the actual values and whether they might bias the comparison.

"Intention to treat analysis"
If a patient is allocated to a surgical treatment group but does not end up with surgery for some reason, which group he/she should be analyzed in?
This is not a simple question.
For example, Feinstein (Principles of Medical Statistics 2002, p. 464) comments:
"The solution to the altered-therapy problem is controversial… The counterargument is that the intention to treat (ITT) approach, although perhaps statistically unbiased, is scientifically improper. With scientific common sense, someone would not be counted as having received surgical treatment if the surgery was not done, nor would patients be counted as having received only medical therapy, if they later had the operation.
Because the ITT controversy has not yet been solved, a possible acceptable compromise is to withdraw the patient as censored when the major therapeutic allocation began…"
Thus the issue of ITT is much more complex than simply to give or not give a point to a trial on the basis of whether ITT approach was reported or not.
Furthermore, Caruso et al. are not consistent when they require ITT analysis in the zinc trials. Caruso et al. do not reject the Karlowski trial [15] although only 190 of the randomized 311 participants

concluded the trial (i.e. 121 dropped, corresponding to 38% of randomized participants). Furthermore, in the Karlowski subgroup analysis 42% of recorded common cold episodes were ignored [19]. That is very far from ITT.

"Methods of statistical analysis" and
"Measurements of probability"
A study can generate valid data, even though the calculations of original authors might be incorrect. For example, in Cochrane reviews it is expected that the Cochrane reviewers extract the data from the study reports, and independent of the original authors, analyze the data themselves. The original statistical analysis is irrelevant to the validity of the collected data.

These comments above do not imply that all the 11 quality items are universally unimportant. They are important in certain contexts, depending on the specific questions and details of the trial. However, the items are heterogeneous and calculating a sum of all the items is inappropriate as a basis for selecting "valid" trials.

****

In their abstract, Caruso et al. write:
"Results: Four studies met all 11 criteria. Three of these studies reported no therapeutic effect from zinc lozenge or nasal spray. One study reported positive results from zinc nasal spray."
http://dx.doi.org/10.1086/520031

Thus, the approach of Caruso et al. was to calculate the sum of all quality points, with a maximum being 11 points, and they accepted only those studies that had the full 11 points.

Dichotomization of study findings to "positive" (statistically significant effect) and "negative" (not statistically significant effect) (3 vs. 1 in the Caruso analysis) is called **vote counting**, which is very simplistic approach to meta-analysis:

Greenland (Modern Epidemiology 2nd edn, 1998 [1], p. 671) comments under title:
"Meta-analysis: Some Methods to Avoid: Qualitative Tally (Vote Counting)"

"*Such a tally can be extremely misleading, even if every single study is methodologically flawless, every study is included, and all the studies are comparable in every relevant respect...*
Mere lack of power might cause most or all of the study results to be reported as null…
The best way to overcome this problem is to apply quantitative methods to pooling of results and base interpretation on estimates, rather than tests of the null."

Given the wide availability of pooling methods in statistical software packages, it is puzzling that Caruso et al. (2007) decided to use vote counting, instead of extracting the actual data and pooling the data. They could have used sensitivity analysis to test the role of methodologically less satisfactory trials on the pooled estimate or regression analysis as suggested by Greenland (see above, p. 672).
Also, the estimates per trial and their confidence intervals are useful. The confidence intervals of "positive" and "negative" trials can be overlapping to such an extent that there is no discrepancy between the trials, even though the trials might be on different sides of the arbitrary P = 0.05 limit.